DOCUMENT RESUME

ED 236 171 TM 830 661

AUTHOR Pecheone, Raymond; Rindone, Douglas A.
TITLE Connecticut Application of Item Response Theory to Large Scale Assessments.
PUB DATE Apr 83
NOTE 10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Policy; *Latent Trait Theory; Measurement Techniques; Policy Formation; Research Methodology; *State Programs; *Test Construction; *Testing Programs
IDENTIFIERS *Connecticut Assessment of Educational Progress; *Connecticut Educational Evaluation Remedial Assist; Connecticut State Department of Education

ABSTRACT

The Connecticut Assessment of Educational Progress (CAEP) is designed to provide legislators and educators with information on the strengths and weaknesses of students in 11 curricula areas. In addition, the Educational Evaluation and Remedial Assistance (EERA) program was developed to assess student proficiency in the basic skills of reading, writing, and mathematics. The role these testing practices play in state policy making is described in this paper. Moreover, the focus of the testing is based on the application of item response theory to test development and the measurement of educational growth. Connecticut is just beginning to use latent trait technology in the development of tests. Advantages and disadvantages of this methodology have emerged and are cited and discussed in this paper. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236171

Connecticut Application of Item Response Theory to Large Scale Assessments

Raymond Pecheone
Connecticut State Department of Education
Bureau of Research, Planning and Evaluation
Post Office Box 2219
Hartford, Connecticut 06145

Douglas A. Rindone
Connecticut State Department of Education
Bureau of Research, Planning and Evaluation
Post Office Box 2219
Hartford, Connecticut 06145

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Paper Presented to the Annual Convention of the National Council on Measurement in Education

Montreal, Canada

Part of a Symposium Presentation Session 10B: Applications of Item Response Theory in Statewide Testing Programs: The State of the States

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. L. Pecheone
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Printed in the U.S.A.

TM 830 661

Connecticuts' Application of Item Response Theory
to Large Scale Assessments

Abstract

The State of Connecticut has embarked on a very ambitious testing strategy. The Connecticut Assessment of Educational Progress (CAEP) is designed to provide legislators and educators with information on the strengths and weaknesses of students in eleven curricula areas. In addition, the Educational Evaluation and Remedial Assistance (EERA) program was developed to assess student proficiency in the basic skills of reading, writing, and mathematics. The role these testing practices play in state policy making is described in this paper. Moreover, the focus of our testing is based on the application of item response theory to test development and the measurement of educational growth. Connecticut is just beginning to use latent trait technology in the development of tests and from our experiences some advantages and disadvantages of this methodology have emerged and are cited and discussed in this paper.

Connecticut Application of Item Response Theory to Large Scale Assessments

Connecticut is rather unique in that it directly ties statewide assessment to a set of measureable objectives in the areas of: motivation to learn, mastery of the basic skills, acquisition of knowledge, competence in life skills, and understanding society's values. Similarly, evaluation of these objectives is backed by legislation (Section 10-4 of the General Statutes) which directs the State Board of Education to "...develop or cause to be developed evaluation and assessment programs designed to measure objectively the adequacy and efficacy of the educational programs offered by public schools." Two important vehicles for achieving these ends are available: (1) the Connecticut Assessment of Educational Progress (CAEP) program and (2) the Education Evaluation and Remedial Assistance (EERA) program.

These two testing programs serve different purposes. CAEP's purpose is to provide legislators and educators with information on the strengths and weaknesses of students in eleven curriculum areas. CAEP is modeled after the National Assessment of Educational Progress (NAEP) and incorporates NAEP items into each CAEP assessment. This approach allows for comparisons to be made at the national, regional and local levels as well as measuring growth over time.

EERA is designed to assess student proficiency in the basic skills of reading, writing, language arts and mathematics, and to identify students in need of remedial assistance in these areas. This paper will focus on the role of Connecticut State Department of Education staff in the actual test development and the use (or non-use) of item response theory for test development in these two assessment areas.

The Assessment, Testing and Evaluation unit which administers these two major assessment programs and also has the responsibility for the evaluation of all compensatory education programs in the state. The units role is broad based, that is, we provide technical advice with respect to these assessments but my unit colleagues have the difficult and rewarding job of implementing the assessment program. In Connecticut, the whole appears to be greater than the sum of its parts. Statewide testing by itself is not sufficient to address the program impact questions which play an important part in the utility and follow through of these testing programs. Assessment in Connecticut is viewed broadly and, by design, we have integrated evaluation with assessment and have expected our personnel to be keenly aware of the interaction of these programs as they impact on school districts. This approach clearly places us in the position of having to be "knowledgeable" consumers of measurement methodology.

Test development, in Connecticut, is not entirely an in-house activity. Certain activities are generally done by an outside contractor through a request for proposal (RFP) process. In addition, the technical aspects of

establishing validity for tests are also included as part of the RFP. People may legitimately ask whether placing test development activities in the hands of outside contractors is wise. This approach should not be misunderstood. Our involvement in test development is not as uninvolved observers but rather as partners in the development process. We have deliberately chosen to expend our department resources on the formative work of insuring the content validity of the test, monitoring the test development procedures, disseminating test reports, and providing technical assistance.

Typically, in a CAEP or EERA assessment, activities are designed to address three phases of development: planning, testing and dissemination. The major activities under each section are:

Planning

1. Establishing an advisory committee;

2. Planning the assessment by choosing the goals, objectives and test items;

3. Field testing all instruments and procedures; and

4. Identifying students to be tested at the designated grade levels.

Testing

1. Testing each of the designated grade levels; and

2. Collecting, scanning, scoring and analyzing the data as they become available

Dissemination

1. Developing a Technical Report, a Summary/Interpretation Report, and a Highlights Brochure; and

2. Providing technical assistance to help local school districts develop policies and programs to meet identified needs.

Several important purposes are served through this approach. In trying to meet the mandate of providing assessment(s) and evaluation information to school districts we are able to deliver quality service in a wide range of measurement areas with a relatively small staff. Secondly, in the test development phase we can put the majority of our eggs in one of the most critical baskets of test development, defining the content of the test. Further, it allows the staff to focus on issues which may broaden the conceptual and methodological boundaries of the test and serves to free up staff time to conduct special studies in these areas of interest. Of course, in any approach, trade offs are made in test development.

Since this symposium is about measurement issues concerning item response theory, let me cite some advantages and disadvantages of using outside contractors for conducting studies involving item response theory.

Advantages

1. Experts can be employed who are knowledgeable about the most current advances in the item response field;

2. The most up-to-date computer technology for developing a new test can be utilized;

3. The credibility of the test development process is enhanced by employing national experts; and

4. This approach is resource efficient in terms of both costs and staff.

Disadvantages

1. Accountability for the test development is not fully in our control;

2. Some aspects of the interpretation and validity of item response theory must be taken on faith. Without an in-house capability to verify the data analysis we are vulnerable to potential mistakes; and

3. Since this is a contractual arrangement, special analyses must be negotiated.

One more aspect of our oversight of the test development process needs to be mentioned. Because validation and verification are not in-house activities, we generally employ an expert in measurement theory to act as an independent advisor to the project. The duties of the advisor are to review the work of the primary contractor and to offer, when appropriate, advice on other measurement strategies which might be more appropriate. In effect, this person provides the verification function which we need as well as acting as a technical advisor to staff.

Connecticut Assessment of Educational Progress (CAEP)

Latent trait theory is just beginning to be applied to the CAEP program. In past years classical methods of test development were solely used (p-values approach). The need for the consideration of latent trait theory came from two forces:

1. State legislation requires measurements of growth over time. State legislators want to know how Connecticut students compare to students nationally, regionally, and locally.

2. The need exists to remain current and update tests as new priorities emerge in curricular areas. The current CAEP assessment illustrates this, the assessment of language arts in reading will be expanded to include additional subtest areas such as listening, writing, literature and computer literacy.

The first application of latent trait theory in CAEP was to compare the performance of the students in grades four, eight and eleven who took the CAEP mathematics test in 1976-77 with students who took the test in 1979-80. Both tests included a common core of items but different students were examined during these assessments. In order to compare the total achievement of students over time, it was necessary to equate the two tests. This requires (a) choosing an appropriate item response model, (b) calibrating the items, and (c) scaling the items so they are on a common scale. Once the items are placed on a common scale, the abilities of the examinees can be estimated and the two groups compared. The one parameter model (Rasch) was selected as the appropriate model and the items were calibrated using the Bical program. A three step procedure was used to complete the comparison analysis. First, testing the adequacy of the fit of the Rasch model to the data was examined. Next the abilities of the examinees with the calibrated items and scaled item parameters were estimated. Finally, the items were placed on a common scale and the mathematics achievement of the two groups were compared with a t-test. The results obtained indicate that the Rasch model fitted the data well. Items which contributed to the lack of fit were deleted from the analysis.

The equating process just described reflects the currently acceptable path that many test developers follow when using the Rasch model. However, because of the need to measure growth in mathematics over time this created some interesting modifications to the Rasch analyses in Connecticut. The maximum likelihood procedure which is the accepted procedure for estimating abilities was not suitable for our purposes because these estimates could not be obtained in the situation where an examinee gets all items correct or no items correct. The standard process is to eliminate these examinees and then do the maximum likelihood estimations. Since different numbers of individuals may be eliminated from the 1976-77 and 1979-80 CAEP sample, a comparison of the students may not be justified. To accommodate for this problem a Bayesian estimation procedure[1] was employed in this situation.

Another modification of the Rasch analyses was made with respect to assessing the fit of the model to the data. While the Rasch fit statistics are useful, they are not completely reliable since their sampling distributions are not fully known. Therefore, if the item parameters after being administered to two samples do not fall on a straight line with a slope 1, the possibility exists that the Rasch model is inappropriate. Due to this property it is important to target items which fall outside ± 2 standard deviation units and delete these items. To accomplish this, the difficulties of the items which are common in the two test forms were calculated and plotted against each

[1] Swaminathan, H and Gifford, J.A. Bayesian estimation in the Rasch model. Journal of Educational Statistics, 1982, in press

other. Note that these plots were done prior to placing these items on a common scale. In general, the number of misfitting items was small which provided further evidence that the Rasch model fit the data well.

Both the analysis of mathematics achievement over time and the Bayesian estimation procedures for the fit analyses were conducted by our consultant Dr. Hariharan Swaminathan of the University of Massachusetts. The application of these methodologies illustrates, in our view, one of the advantages of outside consultation. It is very possible these procedures might not have been employed if we conducted our own analyses. This might be due to lack of expertise, lack of computer technology, and lack of time but, for whatever reason, it seems apparent that we benefit greatly from this kind of collaboration.

Education Evaluation and Remedial Assistance (EERA)

At this time, the EERA assessment is a more standard application of item response theory. The major thrust of our use is in the development of two parallel forms in language arts and in mathematics for the EERA program. In language arts, Test Form B has a total of 36 items in three domains: (1) mechanics, (2) composition and organization skills, and (3) library skills. In mathematics, Test Form B has a total of 65 items in three domains: (1) computation, (2) concepts, and (3) problem solving.

To conduct this horizontal equating study, Test Form B in language arts and mathematics were administered along with a set of imbedded experimental items. Test Form B in language arts was administered with 20 different sets of 6 experimental items. In mathematics, Test Form B was administered along with 20 different sets of 10 experimental items (see Attachment 3). Each of the 20 different sets in language arts and mathematics were administered to approximately 2,000 students. The outcome of this process is that two new test forms will be developed (Test Forms C and D). For both language arts and mathematics, the new test forms are to have the following characteristics:

1. Test Forms C and D will have the same number of items as Test Form B,

2. Test Forms C and D are to be equal in difficulty to each other and to Test Form B at both the domain and total test level, and

3. Test Forms C and D are not to contain any overlapping items.

The Rasch model was used to carry out this equating procedure. The construction of Test Form C and D in language arts and mathematics represents a particular application of the general process of item pool/item bank development. This process has involved three major steps: (a) separate calibration of each test form, (b) equating of all test forms on a common scale, and (c) the selection of experimental items from the item pool to create Test Forms C and D.

The calibration of the items was performed using the Bical program developed by Wright and Mead. Assessing the fit of the data to the model addressed two questions: (1) whether the items fit the model, and (2) whether the subjects fit the model. Misfitting items were flagged if the magnitude of their fit statistic is clearly larger than the fit statistics for the rest of the items on the test. Although this procedure includes the use of some subjective judgement, it was recommended over the mechanical process of comparing the fit statistic to a specific critical value. Items that did not fit the model were handled quite easily. Since the task was to construct new forms, non-fitting items were simply deleted from the item pool. In addition, it is planned that a person fit analyses will also be conducted. Person fit analysis is based on the comparison of the pattern of correct responses a student makes compared with the pattern predicted by the Rasch model. While this analyses is not routinely performed, it may shed some light on anomolies in the data set that will help in the construction of Test Forms C and D.

The second major step is the linking of the items onto a common scale. Recall that each of the anchor items will actually have 20 different anchor items, one for the separate calibration analysis of each test form. To place the items on a common scale, it was recommended that the mean difficulty over the 20 replications be used for estimating the equating constant. After the linking constant was derived, the quality of the linking analysis was also evaluated by means of a regression analysis. The residuals from the regression of the difficulty of the anchor items onto the mean difficulty of the 20 test forms was done to test for the invariance of these estimates. Overall, the development of the item pool for Connecticut in language arts and mathematics was successful.

Finally, the construction of the test forms was quite straightforward. First, all the anchor and experimental items in each domain were rank ordered according to item difficulties. Items are then selected for Test Forms C and D by matching their difficulties to the difficulties of the anchor items which served as a target. To test the adequacy of this matching, the items for Test Forms C and D were identified so that the mean item difficulties within each domain are equivalent for Test Forms B, C and D. In the final analysis this process worked quite well and the development of parallel forms using the Rasch model appears to have been successful.

One final note, since the EERA test will be used for making mastery decisions the reliability of the test around the Connecticut cut scores was also examined. Two reliability estimates were used, Huynh's approximation to Kappa and Livingstone's reliability measure. In general, the reliability of the test forms were quite high.

To summarize, we are just beginning to use item response theory for our test development in Connecticut. We have been quite pleased with our policy to use outside consultants for the actual running of the analyses, although, in the future, we are planning to conduct much more of the verification of the data analyses ourselves. Finally, since we are a state that conducts many types of assessments, it is anticipated that there will be many future applications of latent trait methodology in Connecticut.